

ORACLE

Машинное обучение в базе данных

Oracle

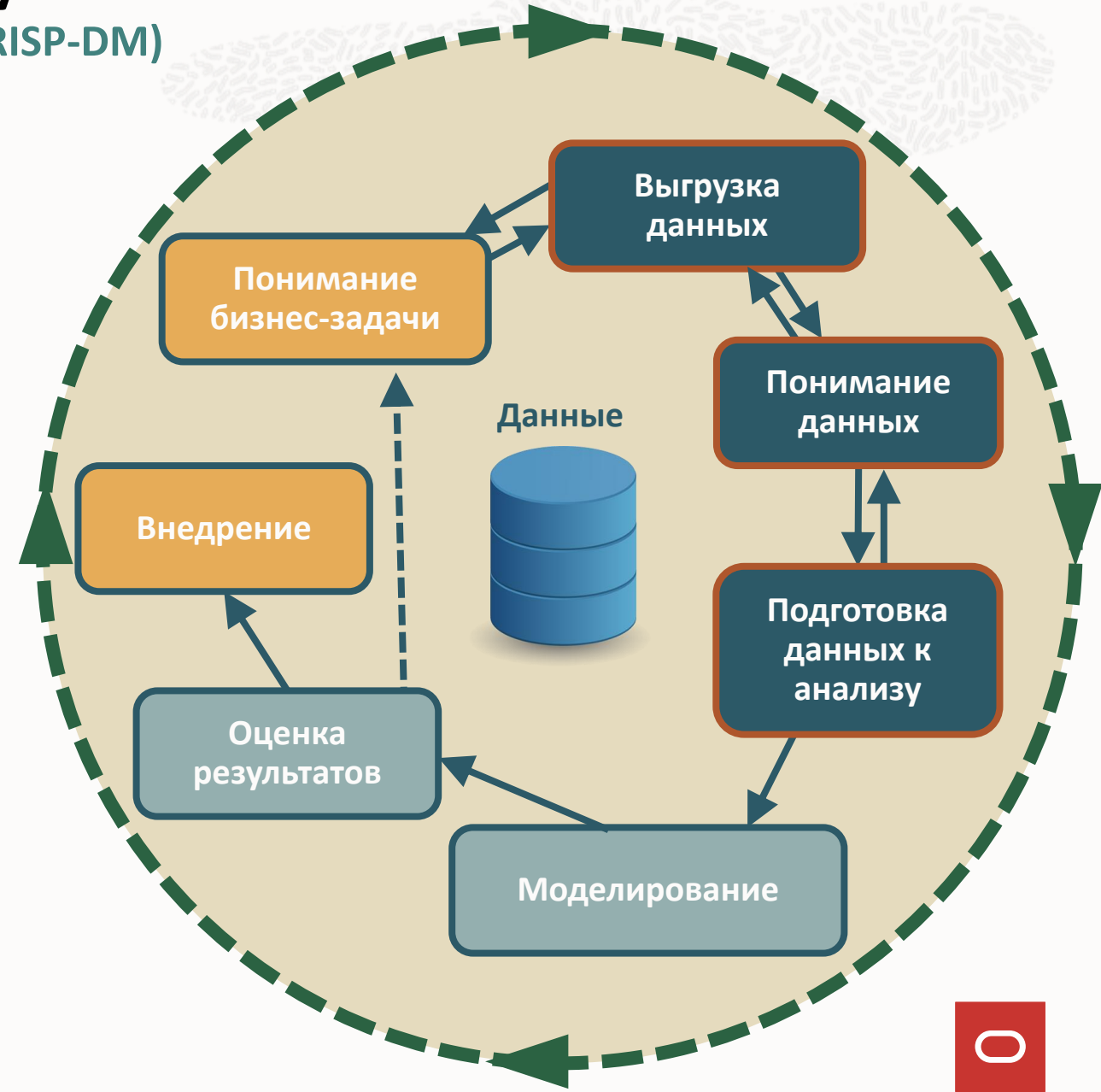
Александр Бернадский

Ведущий консультант

Процесс внедрения машинного обучения

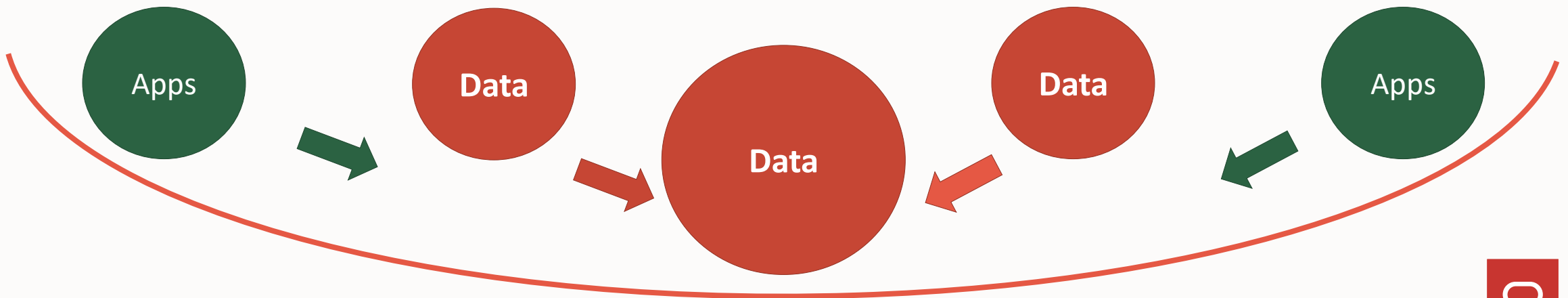
Cross Industry Standard Process for Data Mining (CRISP-DM)

- **Машинное обучение** (ML/Data Science) – это творческая работа с моделями и данными
- Модель сильно **зависит от данных**, и даже небольшое изменение в данных может кардинально изменить подход
- Данные чаще всего объёмны и **плохо описаны**
- Наиболее трудозатратная часть – **управление данными**: выгрузка, понимание, очистка данных, подготовка их для анализа, управление data sets
(Python Libraries, ETL, Data Integration, Data Quality, Autonomous Data Warehouse, BigData, Analytics...)



Гравитация данных

- Есть много задач, для которых нужны данные из всех источников:
 - Корпоративная аналитика
 - Машинное обучение
 - Сквозные бизнес-процессы
 - Маркетинг
 - Биллинг
- Объем данных растёт **экспоненциально**
- Перемещение больших объемов – **медленный** и дорогой **процесс**
- Предпочтительней **избегать хранить данные в разных местах и технологиях**
- Проще **приблизить приложения к данным**, чем данные к приложениям
- Для машинного обучения необходимы вся возможная разная информация
- Лучше обрабатывать данные «на месте их происхождения и хранения»



Универсальная база данных Oracle

Multi-Model and Multi-Workload

Converged Database

Multi-Model

Разные типы данных
(модели и семантика)

Relational, Document, JSON, XML,
OLAP, Spatial, Graph, Object-
Oriented, Text...

Multi-Workload

Разные типы приложений
(нагрузки и технологии)

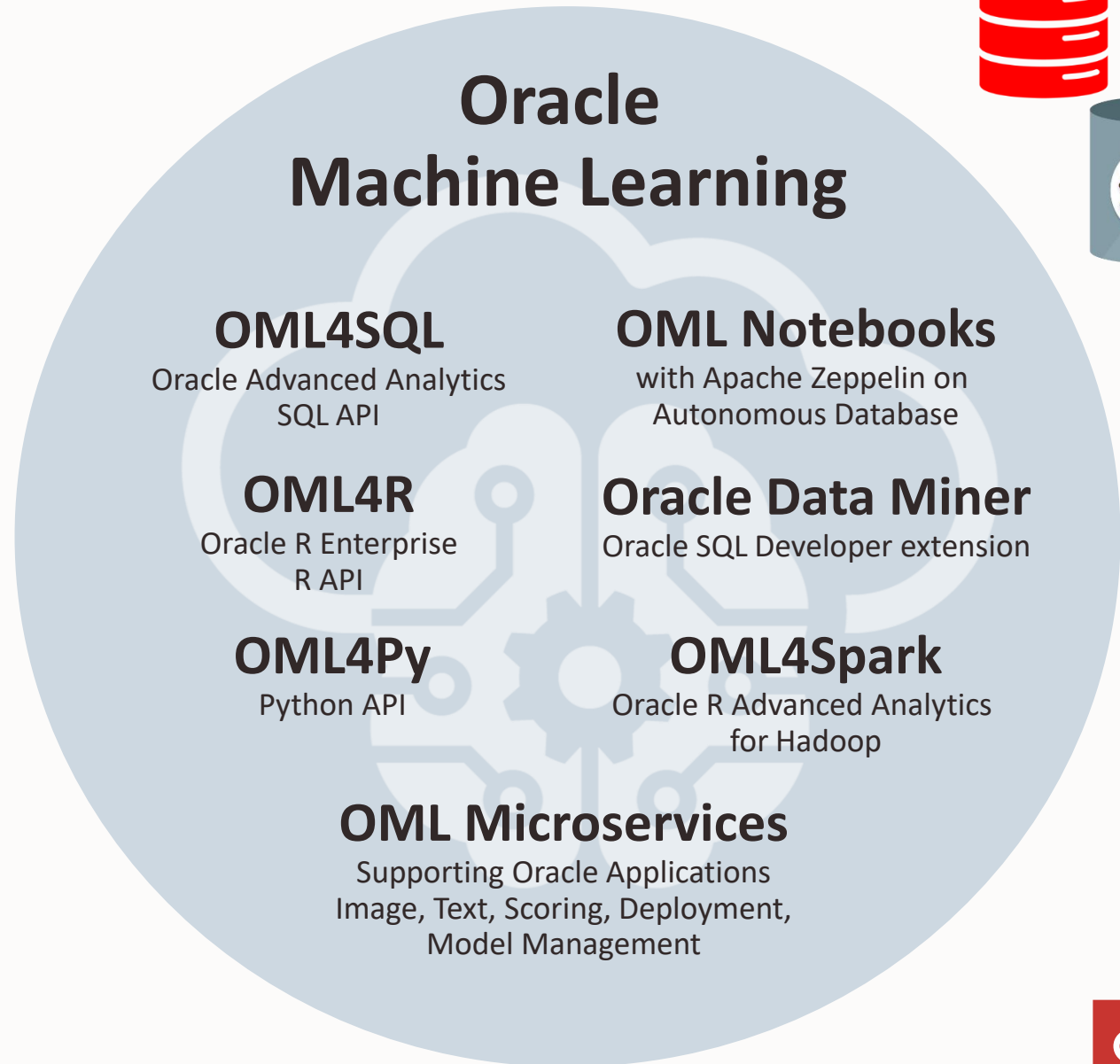
Operational, Analytics, Translytics,
Transactional, IoT, ML, In-Memory, Block-
Chain, HTAP (Hybrid transactional analytic
processing)...

Удобно использовать одну конвергентную базу данных,
которая поддерживает несколько типов данных и рабочих нагрузок



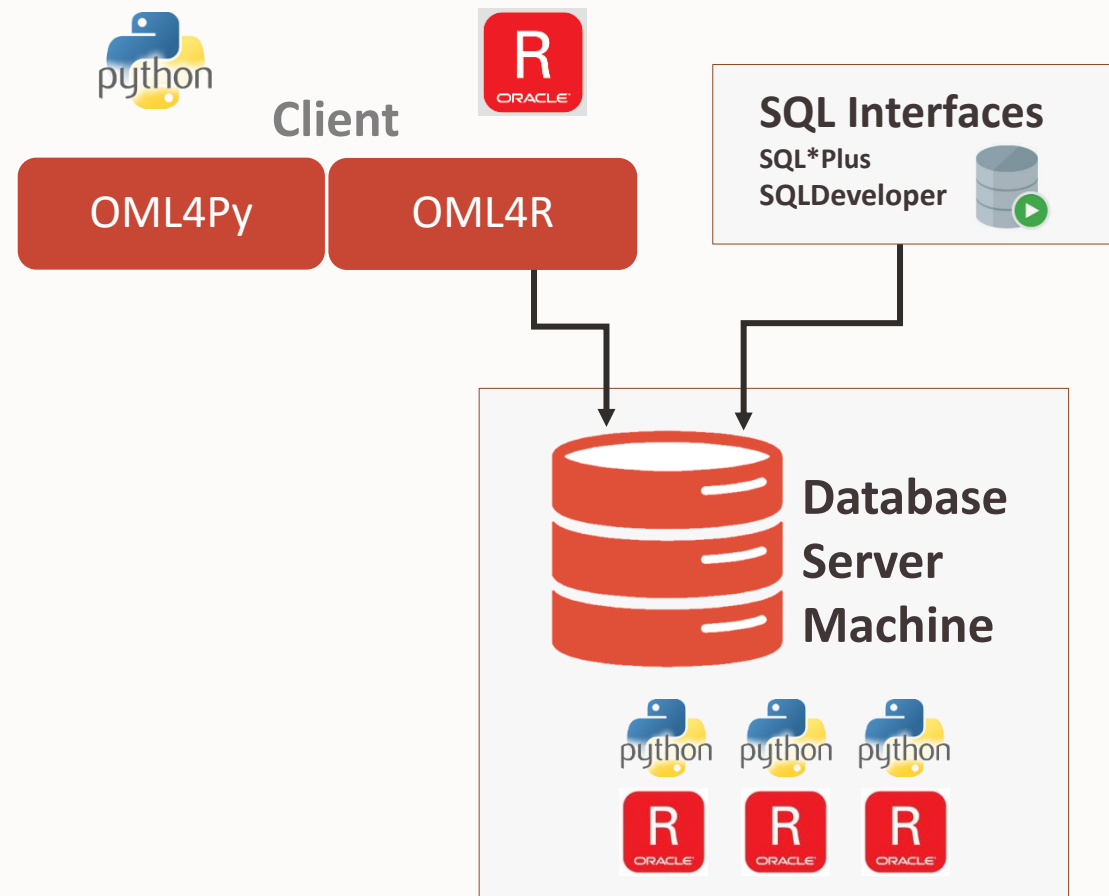
Инновации Oracle в области Machine Learning в базе данных

- Параллельная обработка
- Python, R, PLSQL
- Репозиторий моделей
- AutoML и MLX
- Облачные ноутбуки
- Визуализация данных
- CPU и GPU мощности
- REST и SQL APIs

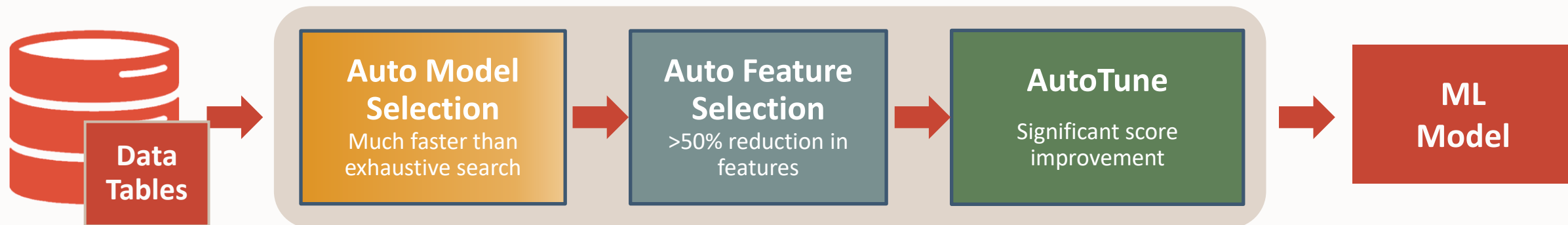


Oracle Machine Learning for Python and R

- СУБД Oracle используется как высокопроизводительная вычислительная среда
- Специальные библиотеки функций для Python/R
- С применением параллелизма и распределенных вычислений
- Объекты и скрипты Python/R управляются СУБД
- Результаты могут быть интегрированы в приложения и дэшборды с помощью SQL



Как работает AutoML в OML4Py



Auto Model Selection

- Определение алгоритма в базе, который обеспечивает самое высокое качество модели
- Поиск наилучшей модели быстрее, чем ручным поиском

Auto Feature Selection

- Уменьшение числа свойств, выделяя наиболее предиктивные
- Улучшение производительности и точности

Auto Tune Hyperparameters

- Значительное улучшение точности моделей
- Замена ручные и поисковых технологий

Основные свойства

- **Минимальное** человеческое участие
- Визуализация выбора и тюнинга модели
- Внедрение в приложения через REST endpoint
- Мониторинг всего процесса с возможностью подстройки/прерывания

Поддерживает методы:

AdaBoostClassifier
DecisionTreeClassifier
ExtraTreesClassifier
KNeighborsClassifier
LGBMClassifier

LinearSVC
LogisticRegression
RandomForestClassifier
SVC
XGBClassifier





OCI Data Science

Совместная работа

UI на основе проектов позволяет командам работать вместе на всех этапах разработки моделей

Open Source

Поддерживает наиболее популярные Open Source инструменты: Python, Tensorflow, Keras, Scikit-learn, MXNet...

Промышленный уровень

Надежность, безопасность и производительность удовлетворяют требованиям крупного бизнеса

CERN. Европейский Центр ядерных исследований



Самая
большая
криогенная
система в мире

15Тб
лог файлов
ежедневно

Используют Oracle Database и
встроенный R для
прогнозирования отказов
оборудования

Цитата



Over the next 10 to 15 years, we expect the total amount of data produced by the experiments on the LHC to increase significantly.



Alberto Di Meglio, head of CERN openlab

NHS BSA

Подразделение национальной службы здравоохранения UK

- Выявлены факты мошенничества на сумму £1.3 млрд
- На 7% сокращение рецептов с антибиотиками
- Миграция аналитического решения с мейнфремов на Oracle существенно сократила IT затраты
- 4000 пользователей, 8000 запросов в день
- План роста до 75000 пользователей на облачной инфраструктуре Oracle



“The NHS sits on billions of data points that have the potential to deliver tremendous value to the wider healthcare system when combined and analyzed effectively.”

Nina Monckton
NHS Business Services Authority

Машинное обучение Для вашего бизнеса

Просто
Быстро
Выгодно





ORACLE

